БИОФИЗИКА, 2022, том 67, № 3, с. 451–455

МОЛЕКУЛЯРНАЯ БИОФИЗИКА

УДК 541.11

ОПТИМИЗАЦИЯ ВЫЧИСЛЕНИЙ В ЗАДАЧЕ СТРУКТУРНОГО МОДЕЛИРОВАНИЯ УРАВНЕНИЯМИ ДЛЯ БИОИНФОРМАТИЧЕСКИХ ПРИЛОЖЕНИЙ

© 2022 г. Г.А. Мещеряков*, **, В. А. Зуев*, А.А. Иголкина*, М.Г. Самсонова*

**Санкт-Петербургский политехнический университет Петра Великого, 195251, Санкт-Петербург, Политехническая ул., 29*

***Институт белка РАН, 142290, Пущино Московской области, Институтская ул., 4*

E-mail: m.samsonova@spbstu.ru

Поступила в редакцию 14.03.2022 г.
После доработки 14.03.2022 г.
Принята к публикации 23.03.2022 г.

Структурное моделирование уравнениями – это метод для анализа линейных взаимодействий между наблюдаемыми и латентными переменными, представленными в виде направленного причинно-следственного графа. Он является популярным инструментом в самых различных областях, от гуманитарных до естественно–научных. За последнее десятилетие данный метод стал особенно интересен в областях, находящихся на стыке с биологией. Однако зачастую в биологических данных нарушается распространенное предположение о независимости наблюдений, что необходимо учитывать на этапе построения математической модели. Кроме того, в такой задаче, как, например, полногеномный поиск ассоциаций, время оптимизации параметров модели является особенно критичным фактором. В данной работе предлагается новая модель метода, а также быстрый способ оценки ее параметров.

Ключевые слова: SEM, structural equation modelling, структурное моделирование уравнениями, semopy, квадратуры гаусса, полногеномный поиск ассоциаций.

DOI: 10.31857/S0006302922030048, **EDN:** ANBULQ

Структурное моделирование уравнениями (Structural Equation Modelling – SEM) – это совокупность подходов к многомерному анализу причинных отношений между наблюдаемыми и латентными переменными. Метод SEM находит применение в широком спектре областей: от психологии и социологии до эконометрики и биологии [1]. Одной из сильных сторон SEM является возможность в явном виде задавать причинно-следственные связи между переменными (или, другими словами, задавать структуру *генеративной модели*), что в случае, если предполагаемые связи соответствуют действительным, увеличивает статистическую мощность модели. Кроме того, зачастую сами причинно-следственные связи являются объектом проверки гипотез, в частности, в гуманитарных исследованиях. Другая сильная сторона SEM – это возможность учитывать в модели латентные переменные. Зачастую, разумное их добавление позволяет более точно описать процесс *генерации* данных и, как следствие, также повысить статистическую мощность модели.

За последние годы возрос интерес к применению SEM в биоинформатических задачах, от анализа воздействия фотосинтеза в период вегетации на сроки старения листьев [2] до исследования поведения генных сетей у больных шизофренией [3] и применения многоцелевой многолокусной модели, использующей моделирование структурных уравнений для описания сложных ассоциаций между однонуклеотидными полиморфизмами и признаками (multi-trait multi-locus Structural Equation Modelling – mtmlSEM) при полногеномном поиске ассоциаций (Genome-Wide Association Studies – GWAS) [4]. Однако, особенно в последнем случае, применение SEM остается ограниченным ввиду отсутствия возможности учесть общую дисперсию между наблюдениями, т. е. отбросить предположение о независимости наблюдений, свойственное большинству линейных моделей (и являющимся одним из условий теоремы

Сокращения: SEM – структурное моделирование уравнениями, GWAS – полногеномный поиск ассоциаций, LMM – смешанная линейная модель, GP – гауссов процесс.

Гаусса–Маркова о независимости ошибок). Взаимозависимость в данных может быть обусловлена либо наличием генетического сродства между образцами и/или их географической близостью. В первом случае, применительно к GWAS, существуют известные подходы на основе смешанной линейной модели (Linear Mixed Model – LMM) [5], а во втором – более общие методы на основе гауссовых процессов (Gaussian Process – GP). Тем не менее, ни LMM, ни GP не обобщены на случай SEM и работают лишь с простейшими линейными моделями.

Такие задачи, как GWAS, требуют высокую скорость работы программ ввиду необходимости обработки больших массивов данных (десятки и сотни тысяч однонуклеотидных полиморфизмов, сотни образцов). В настоящей работе показано, что в общем случае сложность работы алгоритма оптимизации зависит кубически от количества образцов и использование алгоритма в такой форме представляется долгим процессом. Таким образом, целью данной работы являлось: 1) разработать SEM-модель, способную работать с зависимыми данными подобно LMM/GP; 2) добиться меньшей асимптотической сложности, нежели $O(n^3)$.

МАТЕРИАЛЫ И МЕТОДЫ

Была взята модель из программного обеспечения **semopy** [6]:

$$\begin{cases} H = \mathrm{B}H + RG + E, E \sim MN(0, \Psi, I_n) \\ P = \Lambda H + \Pi G + \Delta + U, \Delta \sim MN(0,\ \Theta,\ \mathrm{I_n}), U \sim MN(0, D, J + K) \end{cases}, \quad (1)$$

где P – матрица фенотипов; H – матрица латентных переменных; G – матрица генотипов/однонуклеотидных полиморфизмов; B, Λ, Π, R и Θ, Ψ, D, K – параметризованные матрицы загрузок и ковариаций соответственно (подробнее см. в исходной работе по mtmlSEM [4]), а MN – матрично-нормальное распределение [7]. В отличие от оригинальной модели **semopy** мы, во-первых, перешли от векторной нотации к матричной, а во-вторых, ввели слагаемое U, названное матрицей *случайных эффектов* (англ. *random effects*), контролирующей общую дисперсию между наблюдениями (таким образом, после вычета U из P можно сказать, что наблюдения становятся независимыми). P как сумма линейных трансформаций матрично-нормальных случайных величин также будет матрично-нормальной случайной величиной, и для оценки параметров в матрицах B, Λ, Θ, Ψ, K можно воспользоваться методом максимального правдоподобия.

Опуская прочие выкладки и выразив P из модели (1):

$$P = \Lambda \mathrm{C}(\mathrm{RG} + \mathrm{E}) + \Pi \mathrm{G} + \Delta + \mathrm{U}, \quad (2)$$

где $\mathrm{C} = (\mathrm{I} - \mathrm{B})^{-1}$, посчитав $\mathrm{M} = \mathrm{E}[P]$:

$$\mathrm{M} = (\Lambda \mathrm{CR} + \Pi)G, \quad (3)$$

$\mathrm{V} = \mathrm{cov}[P]$:

$$\mathrm{V} = \mathrm{tr}\{\Sigma\}\ \mathrm{I}_n + \mathrm{tr}\{D\}K, \quad (4)$$

где $\Sigma = \Lambda \mathrm{C} \Psi \mathrm{C}^{\mathrm{T}} \Lambda^{\mathrm{T}} + \Theta$, и $\mathrm{T} = \mathrm{cov}[P^{\mathrm{T}}]$:

$$\mathrm{T} = n\Sigma + \mathrm{tr}\{K\}D, \quad (5)$$

а затем подставив выражения (3)–(5) в логарифм функции плотности матрично-нормального распределения и помножив на –1, получим следующую функцию цели:

$$\mathrm{L}(.) = \mathrm{tr}\{\mathrm{V}^{-1}(\mathrm{Z} - \mathrm{M})^{\mathrm{T}}\mathrm{T}^{-1}(\mathrm{Z} - \mathrm{M})\} + m\ln|\mathrm{V}| + n\ln|\mathrm{T}|, \quad (6)$$

где m – число фенотипов. Минимизация выражения (6) дает оценку параметров модели методом максимального правдоподобия. Однако, как видно из выражения, при каждом вызове целевой функции n необходимо считать V^{-1} (размерность $V - n \cdot n$, где n – число наблюдений), а операция обращения матрицы занимает $O(n^3)$ операций в общем случае.

Прежде всего заметим, что на практике очень важно найти правильную форму матрицы K, так как она определяет поведение слагаемого U. Ясно, что в общем случае K не может быть полно-

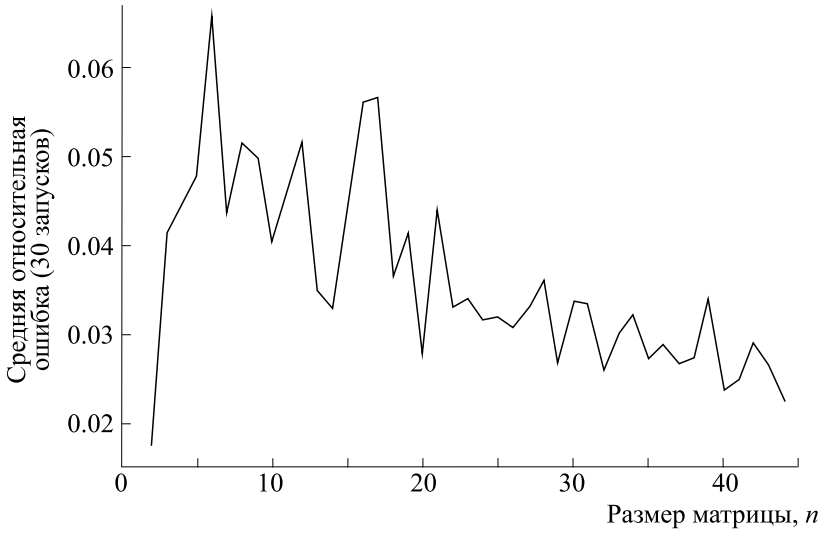

Рис. 1. Зависимость средней относительной точности от размера матрицы n.

стью параметризованной (так как в таком случае число параметров будет порядка n^2, что намного больше, чем число имеющихся наблюдений), и мы вынуждено ограничены в свободе выбора K. Иногда мы можем работать с полностью определенной K. В частности, так делают при необходимости учета генетического родства, когда в наличии имеется матрица генотипов G, а K считается как ковариационная матрица по G [8]. Тогда, используя спектральное разложение K, имеем:

$$K = \mathrm{QSQ}^{\mathrm{T}}. \tag{7}$$

После этого, повернув данные на Q, при подсчете выражения (4) видим, что от K остается только S, и V приобретает диагональную форму, обращение которой занимает O(n) операций. Схожий подход использует инструмент для GWAS FastLMM [5].

Однако зачастую мы не можем знать K точно. Например, если мы хотим учесть сродство, обусловленное географическим положением образцов, то, подобно тому, как это часто делают в GP [9], применяют ядро Матерна для подсчета ковариации между различными географическими точками:

$$cov(i,j) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2}\frac{d_{ij}}{\rho}\right)^{\nu} K_{\nu}\left(\sqrt{2}\frac{d_{ij}}{\rho}\right), \tag{8}$$

где v и ρ – параметры, которые необходимо оценить, а d_{ij} – расстояние между двумя точками. Так, учитывая то, что v и ρ различны на каждом шаге оптимизационного алгоритма, разложение (7) необходимо делать каждый раз, равно как и поворот данных на новое Q, что делает данный подход полностью бессмысленным. К сожалению, в общем случае, невозможно обращать K быстрее чем за O(n^3), но возможно считать «сложные» члены выражения (6) приближенно. В основе вычислительной схемы лежит процедура Гаусса–Ланшоца [10], ранее адаптированная к LMM [11, 12] и находящая применение у GP в рамках программного обеспечения LanczOs Variance Estimates [13]. Процедура Гаусса–Ланшоца позволяет приблизительно оценить квадратичную форму матрицы f(A), где f – любая матричная аналитическая функция, без подсчета f (так, например, $f(\mathrm{A}) = A^{-1}$). Отбрасывая выкладки и детали, укажем, что процедура основана на стохастической оценке следа (так называемый трюк Хатчинсона):

$$\forall x : E[x] = 0, E\left[x^T x\right] = 1 : E\left[x^T f(A) x\right] = tr\{f(A)\} \approx \frac{1}{L}\sum_{i=1}^{L} x_i^T f(A) x_i, \tag{9}$$

и представлении его через интеграл Римана-Стильетса:

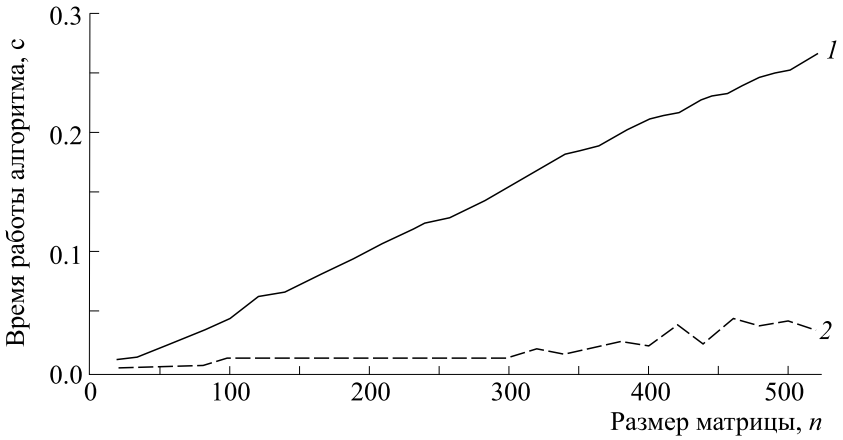

Рис. 2. Зависимость времени работы алгоритма в секундах от размера матрицы *n*: *1* – подсчет по модели (1), *2* – подсчет по предложенному авторами методу.

$$x^T f(A)x = x^T f\left(Q^T \mathrm{S} Q\right)x = x^T Q^T f(s) Q x = \sum_{i=1}^{n} f(s_i)\mu_i^2 = \int_{s\_n}^{s\_1} f(t)\, d\mu(t) \approx \sum_{i=1}^{v} w_i f(b_i). \quad (10)$$

Выражение (10) считается приближенно квадратурами посредством ортогональных полиномов Гаусса–Ланшоца. Чем выше L в выражении (9), тем больше точность алгоритма.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Предложенная модель (1) была успешно реализована и интегрирована в программное обеспечение **semopy** (см. сайт semopy.com). Использованная вычислительная схема дает приемлемую точность (средняя относительная ошибка не превышает 1% уже при L = 3, см. рис. 1) при значительно меньшем времени работы (см. рис. 2). Кроме того, предложенная модель легко расширяется на случай нескольких случайных эффектов с различными K, что дает возможность одновременно учитывать генетическую и географическую родственность образцов данных.

ФИНАНСИРОВАНИЕ РАБОТЫ

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (грант № 18-29-13033).

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Настоящая работа не содержит описания исследований с использованием людей и животных в качестве объектов.

СПИСОК ЛИТЕРАТУРЫ

1. А. А. Иголкина и М. Г. Самсонова, Биофизика **63** (2), 139 (2018).
2. X. Lu and T. F. Keenan, Global Change Biology **28** (9), 3083 (2022).
3. A. A. Igolkina, C. Armoskus, J. R. Newman, et al., Front. Mol. Neurosci. **11**, 00192 (2018).
4. A. A. Igolkina, G. Meshcheryakov, M. V. Gretsova, et al., BMC Genomics **21**, 490 (2020).
5. C. Lippert, J. Listgarten, Y. Liu, et al., Nat. Methods **8**, 833 (2011).
6. A. A. Igolkina and G. Meshcheryakov, Structural Equation Modeling: A Multidisciplinary Journal, **27** (6), 952 (2020).
7. A. K. Gupta and D. K. Nagar, *Matrix Variate Distributions* (Routledge, 1999).
8. J. Goudet, T. Kay, and B. S. Weir, Mol. Ecol. **27** (20), 4121 (2018).
9. C. E. Rasmussen, *Gaussian Processes in Machine Learning* (Springer, Berlin, 2003).
10. S. Ubaru, J. Chen, and Y. Saad, SIAM J. Matrix Analysis and Applications **38**, 1075 (2017).
11. R. Border and S. Becker, BMC Bioinformatics **20**, 411 (2019).
12. Г. А. Мещеряков, в сб. *Тезисы XXI Всерос. конф. молодых ученых по математич. моделированию и информационным технологиям* (2020), сс. 27–28.
13. G. Pleiss, J. Gardner, K. Weinberger, and A. Willson, In *Proc. Int. Conf. Machine Learn*. (2018). DOI: 10.48550/arXiv.1803.06058

Optimization of Computations for Structural Equation Modeling with Applications in Bionformatics

G.A. Meshcheryakov*, **, V.A. Zuev*, A.A. Igolkina*, and M.G. Samsonova*

**Peter the Great St. Petersburg Polytechnic University, Polytekhnicheskaya ul. 29, St. Petersburg, 195251 Russia*

***Institute of Protein Research, Russian Academy of Sciences, Institutskaya ul. 4, Pushchino, Moscow Region, 142290 Russia*

Structural Equation Modeling (SEM) is a technique for analysis of linear relations represented as the causal and correlational relationships between observed and latent variables. SEM is a popular tool in a wide range of fields, from the humanities to the natural sciences. Over the past decade, this method has become especially interesting in areas that are at the interface with biology. However, a common assumption that observations are independent is often violated in biological data; it should be taken into account at the stage of constructing a mathematical model. In addition, in genome-wide association studies, the time of optimization of model parameters is a particularly critical factor. In this paper, we propose a new SEM model, as well as a fast way to estimate its parameters.

Keywords: SEM, structural equation modeling, semopy, Gauss quadrature, genome-wide association studies